Some Statistical Issues in Medicine and Forensics

by

Seymour Geisser*
University of Minnesota

Technical Report No. 564
October 1991

*Research supported in part by NIH Grant GM 25271 and the Lady Davis Trust

Some Statistical Issues in Medicine and Forensics

Seymour Geisser[1,2]
University of Minnesota

1. Introduction

What I intend to do is offer some comments, criticisms and suggestions on a variety of topics. Several are in the area of forensics and litigation and several have to do with Biomedical issues. The particular topics tend to be somewhat disparate so I hope you will bear with me as I make a somewhat disjointed transition from one topic to another. Before I start I will briefly report on an apparently new historical terminology which both peeved and amused me.

A book called "Other Losses" was published in Canada by James Bacque (1989). It received considerable media attention mainly in Canada and Germany since it asserted that close to one million German prisoners of war held by the Americans and French during the World War II period, perished through the deliberate efforts of General Eisenhower to satisfy a personal hatred of Germans (a priori difficult to believe given that Eisenhower's ancestors were German). What piqued my interest was a comment by the N.Y. Times Book reviewer, S.E. Ambrose (1991) who stated that the author's statistical methodology was hopelessly compromised. It appeared that most of the purported "statistical" evidence used to buttress the author's argument that Eisenhower starved the German P.O.W.'s to death depended on missing records, missing bodies, missing numbers, and lastly, missing orders on Eisenhower's part to reduce prisoner rations. The author, James Bacque,

---

[1]President's Invited Address, ASA meetings, Atlanta, Georgia, 1991

[2]Work supported in part by NIH grant GM 25271 and the Lady Davis Trust

neither historian nor statistician but previously a writer of fiction, documents his case by imputing, adjusting and devising incriminating numbers, especially where data are missing to support this extremely unlikely hypothesis. Was this compromised statistical methodology? Another reviewer, A.E. Cowdry (1990), who reviewed the book in a journal of medical history and whose statistical competence exceeded that of the Times reviewer, correctly labeled it as just bad arithmetic, inability to read a table of numbers and a misunderstanding of definitions among other things. Apparently compromised statistical methodology is now social science jargon for faulty arithmetic.

2. Forensics and Litigation

2.1. Scientific Misconduct

At any rate, the kinds of things that statisticians are called upon for advice or consultation are extraordinarily diverse. Not too long ago I received a telephone call from an employee of a Federal agency, who after introducing himself said "I understand that you are an expert in the falsifying of data". What was that, I responded? He thought I misheard so he repeated himself. Now as you know, ancient Gaul during the time of Caesar was divided into three parts, so too is modern gall the usual bladder and stones, and the third that appeared to be reflected here - unmitigated.

This sort of view of what a statistician does is not restricted to that of the average citizen but often extends to the highest reaches of our government. Witness the statement made by the Senate Majority leader early this year when commenting on some economic data he was shown, whose particular selection he did not appreciate. He remarked that it was "an obvious statistician's trick". Not an economist's trick nor an

econometrician's trick nor heaven forbid a politician's trick but a statistician's trick. So knavish does our discipline appear even to educated outsiders.

Now back to the phone call. When I seemed offended at the remark made by the man who called - he said no, you misunderstood, what I meant was that as a statistician, you are expert in data falsification. It really didn't sound any better this time - but I asked him somewhat warily what it was he wanted of me. It turned out what he really wanted was someone to look at a Biomedical experimenter's notebooks and published papers and to investigate an allegation that this scientific worker submitted falsified data in order to obtain federal grants. If the evidence indicated this, there would be interest in prosecuting the individual to recover the funds and damages.

In looking over the material it was clear that this experimenter didn't use any esoteric or elegant statisticians' tricks at all. Apparently in reporting data if a mean of a set of quadruplicate measurements did not suit his purpose, he arbitrarily omitted one value in his publications, either the largest or the smallest or sometimes retaining only 2 of them, so that the resulting mean better supported his hypotheses - why not, isn't that what statistics was for, to verify an experimenter's hypotheses?

Among other things, his graphs reflected not his collected data but apparently created data that again demonstrated support for his hypotheses. When deleting an observation he sometimes retained the original sample variance. Either this was out of sheer laziness or his statistics course was so unsophisticated as not to realize that for the value of a sample variance, that deletes the largest or smallest of 4 observations, to remain unchanged requires the omitted value to be a unique function of the other 3 observations. Perhaps he may have had a more sophisticated course that included the fact that the sample mean and sample variance from a normal

population were independent - from which he may have inferred that there was no need to recompute the sample variance. Who knows? What else might one conclude from all of this other than gross incompetence or the more obvious scientific misconduct? At any rate, alleged scientific fraud involving statistical manipulations may not always be as easy to demonstrate, unless the original experimental data sheets have been sequestered, as it was here, by a distrustful laboratory assistant.

Although detecting scientific misconduct or data falsification appears to be a new venue for the statistical enterprise, I need only remind you of Fisher's view of some of Mendel's genetic data, wherein he attributed the extraordinary fit to an overzealous assistant who in Fisher's words "cooked the data" because he knew what was required.

2.2. DNA Fingerprinting

Testimony in court based on the use or abuse of a relatively new forensic tool, under the rubric of "DNA fingerprinting", is being heard with increasing frequency in criminal and paternity trials throughout this country.

"DNA fingerprinting" or pattern matching is of potentially enormous value in problems of individual identification. Its power derives from the fact that the locus on a chromosome that is probed has a multitude of presumably discrete expressions or alleles (estimated at anywhere from 30 to 2000 or more), which after much technical manipulation, namely agarose submarine gel electrophoresis and Southern blotting, are transformed into two band weights measured in DNA base pairs. The bands arise one from each parent but one cannot generally ascertain which band is from which parent.

Several probes often on different chromosomes yield a profile of band values that constitutes the so-called DNA fingerprint. These profiles can be obtained from a variety of human tissue. It is also claimed that for every individual, variation of the same probe between tissues, be it blood, semen, hair, etc., is virtually the same as the variation of repetitions of the probe within any tissue. This increases its potential use by orders of "forensic magnitude" over the usual fingerprint data. Its value rests in its capacity in the preponderance of cases to demonstrate the incompatibility of a forensic sample with that of a victim or suspect depending on circumstances. These are the ones that don't come to court. The ones that do are those claimed to be putative matches.

What is the problem then? First is the fact that the technical procedure is unable to resolve the band values of proximal alleles, thus requiring inferential statistical methods. This in itself need not be a problem, but the use of statistical methods by geneticists, molecular biologists, and biochemists in defining a match between two samples may be. The second major problem involves determining the relative frequency of an individual's pattern on a few probes in an appropriate population, since populations appear to differ in the distribution of band weights. The latter became an issue when infinitesimally small values - as low as one in many trillions - were assigned to this probability or relative frequency in specific instances. This was particularly disconcerting as the size of populations that were used was generally between 100 and 400 for any probe. The ability to achieve such small estimates rested mainly on the assumptions that not only were the probes statistically independent but so were the bands within each probe. These assumptions enabled the use of what geneticists term the product rule (i.e., multiplication of relative frequencies) for the 2 bands within the 4 probes

generally used. Within a probe the "Hardy-Weinberg Law" was invoked and between probes, linkage equilibrium. These genetical concepts are synonymous with statistical independence. Geneticists affirm that the Hardy-Weinberg law is the culmination of a sufficient number of generations in a population in which a large enough proportion in each generation are randomly mating.

The first problem is to decide whether a forensic sample generally obtained at the scene of the crime (which may be somewhat degraded) is a match for a freshly collected sample from a suspect. Various labs use various criteria - after visually checking to see if the two samples are similar or not. If presumed similar it is numerically checked by determining whether at each probe the two sample values are within a given percent of one another; the percent will vary with the laboratory. The distribution of the band weights in a probe reflecting the discrete alleles are considered to be quasi continuous because of their large number and the fact that it is not possible to resolve alleles that are not far apart.

Studies on replicates on the same individual indicate the adequacy of the normal distribution, but with a measurement error that depends on the bandweight of the allele so that the estimated standard deviation is somewhere between .06% and .09% of the bandweight. They also indicate that the larger the bandweight, the larger the percent error in the range studied. To further complicate matters, there is no known distributional pattern of how the alleles vary on a given probe. And there are obvious distributional differences from probe to probe. Coupled with this is the inability to differentiate the bandweights as to maternal and paternal which would not be a problem under approximate bivariate normal theory for the

distribution of the bands in a population. However, this is clearly an inappropriate assumption except if an adequate transformation can be found.

If a match is declared then the sample is compared with a data base that purports to be a sample from the same population as the suspect as an estimate of a random match. The data base generally consists of a size of 100-400 individuals who have been measured on these probes to determine the relative frequency of individuals in that population considered to have the same profile as the suspect. These data bases are often incomplete i.e. some individuals on some probes but not on others and only a fraction of the total on all of them.

Further, the sampling of populations is not only not random but not even haphazard - rather it is catch as catch can, with the necessity that sometimes the catch should be canned as quite often the same individual may appear more than once in a sample. In a recent FBI sample of about 200 men submitted in a Minneapolis court case, it was found by the defense that 8 individuals had been duplicated and 1 triplicated. Further a data base was obtained where an appreciable percentage of individuals had 3 and sometimes 4 bands. No apparent error in the technique could be found to account for this. Some wag suggested the possibility of multiple fathers.

It is of some interest to note that while the standard deviation has been variously estimated as being between .6 - .9% of the band weight, the FBI (Budowle et al. 1991) with some exceptions, will declare a match between two samples if the bands are within 2.5% of the average of the 2, or they can differ by as much as 5%. Note that this is a tolerance of about 4 standard deviations of the difference i.e. things that are as far apart as 4 s.d. will be considered a match. A match then is yes or no without indicating its probability or

likelihood i.e., just within 4 s.d. is considered a match and just beyond is not considered a match.

What is used then is an extraordinarily wide net to declare a match and then a potentially biased data base, usually collected for entirely different reasons and often from geographically distant populations to estimate the profile characteristics of a local population. Different laboratories use different methods to estimate the relative frequency of individuals that have the same profile as the one in question. They tend to neglect sampling error and more importantly assume statistical independence without proper validation. In fact it has been shown that on several probes this is an inappropriate assumption. This has been pointed out by the use of a chi-squared test created from the joint quantiles into an upper triangular contingency table with cells consisting of the number of pairs of observations, both of whose members are in the appropriate bivariate quantiles.

The meager proficiency testing done is on fresh samples rather than on somewhat degraded forensic samples. These tests, scant though they are, indicate that there are about 2% false positive and 2% false negative rates in the matching procedure. These values are incompatible with the infinitesimally small probability estimates of a random match in a data base that are presented in court. Recently this was pointed out in a trial by a defense witness. In order to attempt to nullify the evidence inherent in proficiency testing the prosecutor deliberately mispronounced the name of the witness, after correctly pronouncing it many times. When the witness corrected him as the prosecutor expected he would, the prosecutor retorted that because he had mispronounced the name once in 20 times could anyone believe he would ever do it again! So much for analogy and error rates.

At any rate, with all these problems it is unconscionable that the three laboratories which provide almost all of these analyses have not yet seen fit to employ statistical help in improving their procedures. Some work in the statistical direction using likelihood ratios is discussed in Berry (1991), but the major improvements should come from a potential elimination of the error in identifying an allele by the introduction of the new polymerase chain reaction (PCR) technology and obtaining larger and better samples from appropriate populations.

2.3. Increased Risk and Causal Analysis

In Israel about 300,000 children per year are subjected to DPT (Diphtheria, Pertussis, Typhoid) immunization. There is an accepted risk of encephalopathy leading to irreversible brain damage (IBD) of about 1 case in 310,000 for those given the DPT vaccine, which is slightly higher than among non-vaccinated children. Parents are informed of this risk and accept it in terms of the potential benefits of the vaccine. A case that came up there for litigation was described by Aitken (1991). In one such year there were 4 cases and the parents of one of the victims sued the Ministry of Health, claiming that year's vaccine given to their child was defective. Each side engaged a statistician only to determine the answer to the question of whether there was an increased risk in that year. Note that this bears only on the predictive causal proposition could or will the administration of the new pertussis vaccine increase IBD over the older one but not on the retrodictive proposition (did it) that the vaccine caused IBD in this case. Even if it were virtually certain that the vaccine could cause an increase of IBD, it might be virtually certain retrodictively that it did not do so in this case.

In considering the difference between prediction and retrodiction there is the story of the man who visits a fortune teller. She looks into her crystal ball and tells him that he will lose his job, his wife will leave him and he will have a terrible auto accident. But he tells her, all of that happened to me last week. She looks again into her crystal ball, picks it up, shakes it vigorously and holds it up to her ear and says, the damned thing stopped.

In order to answer the indicated question, both statisticians agreed on the use of a Poisson approximation to the binomial data and the calculations made were based on the assumptions they were given. Hence they agreed to the following calculations. If X = # of cases out of 300,293 vaccinations, then

$$\Pr(X=x \mid H_0) = e^{-.97}/x!$$

x	$\Pr(X=x \mid H_0)$	$\Pr[X \geq x \mid H_0]$
0	.368	1.000
1	.368	.632
2	.178	.253
3	.058	.075
4	.014	.017
5 or more	.003	.003

Both statisticians calculated as a test of the null hypothesis that the risk was 1/310,000,

$$\Pr[X \geq 4 \mid H_0] \doteq .02.$$

The question in dispute was whether this was strong evidence of an increased risk or not. One statistician said that since it did not reach the .01 level

though it exceeded the .05 level the evidence was minimal but not strong evidence of an increased risk. The other argued that it was sufficient. The judge admonished them that he was the one to decide on minimal or sufficient.

It appears that the statisticians were not asked the critical question. That question should have been concerned with the chance that this child's encephalopathy was caused by that year's vaccine. This is a much more difficult question than to assess whether there was an increase in the damage rate. Naively one might go about it as follows: If the datum were 2 cases that year, then the P-value would be .253 and both statisticians would have concurred that there was insufficient evidence to deny the damage rate of 1 in 310,000. One then might argue that if this were the case that one would allow only 2 of the 4 to be attributable to the vaccine. Given the absence of other information one could regard the cases as exchangeable so the chance that this child's condition was caused by the increased damage rate was 1/2. If a P-value of .075 were allowed by both statisticians to be insufficient to reject the H_0 so that 3 cases would still be admissible in denying an increase in the damage rate then the chance would be reduced to 1/4. On the other hand if only 1 was attributable to the usual risk, then probability would be 3/4. However such reasoning might appeal to a judge, what one really needs is to calculate the probability that the vaccine caused IBD in this particular child after marshalling all the relevant evidence.

How was the case resolved? It was settled out of court, as most such cases are, not necessarily in the interest of "scientific truth", whatever that might be in this instance, but for the sake of perhaps higher truths, prudence and economics. But better still, it resulted in a sensible decision by the Ministry to pay compensation for all cases of IBD where a vaccine was

administered since the benefits to the nation as a whole far outweighed the risks. Potential payment could also spur investment in research for safer vaccines.

Among the circumstances of this particular case is the fact that since it is known that the DPT vaccine can cause encephalopathy, it was assumed that it did here and the only question apparently at issue was whether there was a higher risk with the vaccine used during the particular year of this child's vaccination. The arguments in court were restricted mainly to whether a .02 P-value was sufficiently small to establish the claim. Discussions out of court turned on whether a Bayesian analysis involving computation of the posterior odds of increased risk might be more informative than a P-value and the reluctance of both statisticians, well conversant with and often users of the Bayesian approach, to implement it here because they believed standard methods were more widely understood and acceptable.

However, the basic and important question was left unresolved both in a legal or a scientific sense. That is, given an adverse reaction in an individual what is the chance that it was caused by a particular drug, vaccine, therapy or whatever?

Now one promising development of the use of probabilistic methods in causality assessment has been underway for the last few years with special reference to adverse reactions to drugs in clinical practice, Lane (1989). However the methodology is general enough so that it can be applied to many other areas. The method involves a panel of subject experts and a statistician to guide the proceedings which results in recommending a decision based on the expert's retrodiction about the probability of what occurred. These procedures can be guided by one or more or even all of the following principles of the 3 reigning schools of Statistics:

Fisher-Barnard: Separate the relevant information from the irrelevant information.

De Finetti-Savage: Analysis should be logically consistent.

Neyman-Pearson: Minimize the proportion of incorrect decisions.

Here is a rather simple but instructive instance of an analysis of a case of encephalopathy in Canada, Lane (1991). An 8-week male infant who at the time he received his DPT vaccine had mild coryza (inflammation of the mucous membranes of the nose giving rise to sneezing and discharge of mucous). At 12 hours post immunization the child was fevered and 100 hours later developed encephalopathy with decreased state of consciousness and a fever of about 103°F. A radiograph of the chest suggested the possibility of viral pneumonia. The two possibilities that were not ruled out for the encephalopathy were the vaccine and the virus and an attempt was made to calculate the relative odds. Hence using Bayes' theorem,

$$\frac{Pr(V\rightarrow E \mid B,C)}{Pr(v\rightarrow E \mid B,C)} = \frac{Pr(V\rightarrow E \mid B)}{Pr(v\rightarrow E \mid B)} \frac{Pr(C \mid V\rightarrow E,B)}{Pr(C \mid v\rightarrow E,B)}$$

where B = Background, V = vaccine, v = virus, E = Encephalopathy, C = case information and the arrow implies causation. Data were obtained on the incidence of viral encephalopathy not preceded by vaccination in children < 1 year old to be 3.2 per million.

From another study, they estimated an incidence of vaccine induced encephalopathy to be 8.9 per million in the first week following immunization, hence

$$\frac{Pr(V\rightarrow E \mid B)}{Pr(v\rightarrow E \mid B)} = \frac{8.9}{3.2} = 2.8.$$

However, viral encephalopathy is twice as likely to occur in summer than other times i.e. summer incidence is 6.4 per million which is differentially diagnostic, there being no plausible mechanism for causation by the vaccine to exhibit seasonal variation. Now this case occurred during the summer and since viral encephalopathy is twice as likely to occur in summer,

$$\frac{Pr(C_S \mid V\rightarrow E,B)}{Pr(C_S \mid v\rightarrow E,B)} = \frac{1}{2}.$$

Further, on the basis of expert opinion and some mechanistic modeling probabilities of .8, .15 and .05 were assigned to the probability of vaccine-induced encephalopathy for the days 1, 2, 3 following vaccination. For viral encephalopathy each day of a reference week was equally likely, i.e. 1/7 since the onset was on the third day

$$\frac{Pr(C_T \mid V\rightarrow E,B)}{Pr(C_T \mid v\rightarrow E,B)} = \frac{.05}{1/7} = .35.$$

The panel also ascertained that on an average there are 5 viral illnesses during a child's first year, each lasting a week. They also estimated that 10% of children would have a concurrent viral syndrome. Further, the relevant literature indicates that 1/4 of children with presumed viral encephalopathy have a history of antecedent viral syndrome. Thus,

$$\frac{\Pr(C_H \mid V\rightarrow E,B)}{\Pr(C_H \mid v\rightarrow E,B)} = \frac{1/10}{1/4} = .4$$

hence

$$\frac{\Pr[C \mid V\rightarrow E,B]}{\Pr[C \mid v\rightarrow E,B]} = .5 \times .35 \times .4$$

$$\frac{\Pr(V\rightarrow E \mid B,C)}{\Pr(v\rightarrow E \mid B,C)} = 2.8 \times .5 \times .35 \times .4 = .19$$

which now favors a virus induced encephalopathy by about 5 to 1.

A causal analysis of this sort would be more direct and useful than just determining an increased damage rate and obviously more informative to a court than an expert merely stating his opinion as to whether or not an event was caused by a particular agent.

3. Medicine

3.1. Mass Screening

An area involving public policy and health in which statisticians can play an important role is in the planning of mass screening tests. Let me say that these tests are certainly going to come. If mass screening of physicians, dentists and other health care personnel for HIV (Human Immunodeficiency Virus) is a strong possibility, then patients are not far behind if not already done surreptitiously. Screening for venereal diseases prior to marriage has a long history. In my own state there has been a movement towards mass screening for AIDS. Some countries and many states in the U.S. have begun

mass screening of pregnant women for HBV (Hepatitis B Virus) carriers. Infants with less developed immune systems are much more susceptible than adults to the disease which can cause incurable liver cancer later in life.

Since testing for any characteristic is not an inexpensive commodity when conducted on large populations, statisticians should play an important role in the planning of mass screening programs so that testing may be conducted in an optimal or near optimal manner in terms of costs and benefits with appropriate balancing of societal concerns and individual rights. This should not be left solely to physicians and ethicists. For example in the case of HIV, there are at least two tests available for screening purposes - the ELISA and the Western Blot, the first being an order of magnitude less costly than the other but yielding higher false positive and false negative rates and sometimes yielding indeterminate results. Hence whether one needs to choose between the two diagnostic tests, or alternatively start with one and conditional on that result use or not use the other or to use them both sequentially or simultaneously involves a number of statistical questions for efficient implementation.

An optimal procedure will depend on several factors including the prevalence of the condition, the probabilities of false positives and false negatives, the loss in making incorrect decisions, the relative costs of the tests and the costs of their administration. All of these need to be carefully considered and estimated in the context of a given population and application. Even if the cost per individual may vary only slightly from one rule of test administration to another, it is obvious that when these differences are multiplied by hundreds of thousands or even millions, one may be speaking in the late Sen. Dirksen's famous phrase, of "real money".

Consider a mass screening program that has the possibility of using two binary tests say for detecting a condition denoted as C for its presence and $\bar{C}$ for its absence as proposed by Geisser and Johnson (1991). We can denote the outcome T_i and $\bar{T}_i$ respectively as positive and negative on test i = 1,2. These tests can be used in a variety of ways and we will list 8 decision rules out of a larger possible number. Rules other than these 8 in Table 1 will be inadmissible under sensible ranking assumptions that are made on the sensitivities and specificities of the tests and the losses for making incorrect decisions.

Table 1

Rule	Decision Rule T (Assert C if)	Notation
R_1	Test 1 is positive	T_1
R_2	Test 2 is positive	T_2
R_3	Both tests positive (simultaneous tests)	(T_1T_2)
R_4	Either test positive (simultaneous tests)	$(T_1 \cup T_2)$
R_5	Both tests positive (sequential tests)	T_1T_2
R_6	Both tests positive (sequential tests)	T_2T_1
R_7	Either test positive (sequential tests)	$T_1 \cup \bar{T}_1T_2$
R_8	Either test positive (sequential tests)	$T_2 \cup \bar{T}_2T_1$

For decision rule R_8, say, we assert the presence of C if test two was administered first and was positive or if test two was negative and test one was subsequently positive.

We define our loss function in Table 2.

Table 2

		True State: C	True State: $\bar{C}$
Decision rule	T	ℓ_{TC}	$\ell_{T\bar{C}}$
Outcome	$\bar{T}$	$\ell_{\bar{T}C}$	$\ell_{\bar{T}\bar{C}}$

For example, the cost of a positive decision when C is present is ℓ_{TC}.

When C denotes, say, the virus for AIDS, there are two points of view that one can take regarding this cost function; one from the perspective of society and the other from that of the individual. From a societal standpoint, it could be argued that

$$\ell_{TC} \leq \ell_{\bar{T}\bar{C}} < \ell_{T\bar{C}} \leq \ell_{\bar{T}C},$$

for example, if one were accepting blood from screened donors. From an individual's point of view a public expression of being positive for HIV when in fact he is not might be considered worse than saying he is negative when in fact he is positive given the stigma that certain people attach to the disease. On the other hand, even from an individual prospective given treatment for a condition that does not exist certainly may not seem worse than not being given treatment for a condition that does exist. In general, it may only be reasonable to assume that the costs of making a correct decision are less than the costs of making an incorrect decision.

Conditional probabilities for the various outcomes of the two tests are introduced in Table 3:

Table 3

	C		$\bar{C}$	
	T_2	$\bar{T}_2$	T_2	$\bar{T}_2$
T_1	η_{11}	η_{10}	θ_{11}	θ_{10}
$\bar{T}_1$	η_{01}	η_{00}	θ_{01}	θ_{00}

Thus $Pr(T_1,T_2 \mid C) = \eta_{11}$, $Pr(\bar{T}_1,T_2 \mid \bar{C}) = \theta_{01}$ etc. We define $Pr(C) = \pi$ as the prevalence of C. For each of the eight joint tests in Table 1, define the sensitivity and specificity as

$$\eta_i = Pr(T \mid C),\ \theta_i = Pr(\bar{T} \mid \bar{C})\ \ i=1,...,8.$$

For decision rule R_i, the expected loss is

$$E(Loss \mid R_i) = \pi\eta_i[\ell_{TC}-\ell_{\bar{T}C}] + (1-\pi)\theta_i[\ell_{\bar{T}\bar{C}}-\ell_{T\bar{C}}] + \pi\ell_{\bar{T}C} + (1-\pi)\ell_{T\bar{C}}\,. \qquad (3.1)$$

The values of $Pr(T \mid C)$ and $Pr(\bar{T} \mid \bar{C})$ for each of the first 4 decision rules are presented in Table 4.

Table 4

Rule #	Pr(T | C) (sensitivity)	$Pr(\bar{T} \mid \bar{C})$ (specificity)
R_1	$\eta_1 = \eta_{11} + \eta_{10}$	$\theta_1 = \theta_{00} + \theta_{01}$
R_2	$\eta_2 = \eta_{11} + \eta_{01}$	$\theta_2 = \theta_{00} + \theta_{10}$
R_3, R_5, R_6	$\eta_3 = \eta_{11}$	$\theta_3 = \theta_{00} + \theta_{01} + \theta_{10}$
R_4, R_7, R_8	$\eta_4 = \eta_{11} + \eta_{10} + \eta_{01}$	$\theta_4 = \theta_{00}$

The assumptions that lead to consideration of only these 8 rules are

$$\max\{\ell_{TC}, \ell_{\bar{T}\bar{C}}\} < \min\{\ell_{T\bar{C}}, \ell_{\bar{T}C}\},$$
$$\eta_{11} > \eta_{ij} > \eta_{00} \text{ and } \theta_{00} > \theta_{ij} > \theta_{11} \text{ for } i \neq j = 0,1. \tag{3.2}$$

For a mass screening program, such as contemplated for certain diseases, it would be desirable to use an optimal rule. Defining $k = (\ell_{\bar{T}\bar{C}} - \ell_{T\bar{C}}) / (\ell_{\bar{T}\bar{C}} - \ell_{T\bar{C}} + \ell_{TC} - \ell_{\bar{T}C})$, it can be shown that the following necessary and sufficient conditions hold:

$$\begin{aligned} &R_1 \text{ is optimal} \Leftrightarrow Pr(C \mid \bar{T}_1 T_2) < k < Pr(C \mid T_1 \bar{T}_2). \\ &R_2 \text{ is optimal} \Leftrightarrow Pr(C \mid T_1 \bar{T}_2) < k < Pr(C \mid \bar{T}_1 T_2). \\ &R_3 \text{ is optimal} \Leftrightarrow Pr(C \mid T_1 \bar{T}_2) < k \text{ and } Pr(C \mid \bar{T}_1 T_2) < k. \\ &R_4 \text{ is optimal} \Leftrightarrow Pr(C \mid T_1 \bar{T}_2) > k \text{ and } Pr(C \mid \bar{T}_1 T_2) > k. \end{aligned} \tag{3.3}$$

where

$$P(C \mid \bar{T}_1 T_2) = \frac{\pi\eta_{10}}{\pi\eta_{10} + (1-\pi)\theta_{10}}, \tag{3.4a}$$

$$P(C \mid T_1\bar{T}_2) = \frac{\pi\eta_{01}}{\pi\eta_{01}+(1-\pi)\theta_{01}}. \qquad (3.4b)$$

In any large scale screening program costs of administering the tests will be of considerable concern. For example, with regard to AIDS, the ELISA test is an order of magnitude less expensive than the Western Blot. Also, there may be a differential in the costs of a simultaneous administration of both tests in contrast to their sequential administration. Among other things, this may result from having to store the sample until the result from the first test is obtained, or asking a testee to return for testing. Once all of the actual testing costs are carefully ascertained, their incorporation into a complete decision analysis can be made without much difficulty. The major problems are in assessing, in some reasonable way, the original losses on a comparable monetary scale with the actual expense of testing.

Let K_i be the cost of administering test i alone, and let K_{ij} be the cost of administering test i followed by administering test j. Let $K_{(12)}$ be the cost of administering both tests simultaneously. Clearly, it is reasonable to assume that $K_{ij} \geq K_{(12)} \geq \max(K_1,K_2)$. The expected cost under decision rule i is easily calculated and values are given in Table 5.

Table 5

Decision Rule	E(Cost)
R_1	K_1
R_2	K_2
R_3	$K_{(12)}$
R_4	$K_{(12)}$
R_5	$K_1+(K_{12}-K_1)Pr(T_1)$
R_6	$K_2+(K_{21}-K_2)Pr(T_2)$
R_7	$K_1+(K_{12}-K_1)Pr(\bar{T}_1)$
R_8	$K_2+(K_{21}-K_2)Pr(\bar{T}_2)$

The above probabilities in (3.4) are conditional on the parameters, π, θ_i and η_i, i=1,...,8. The costs in Table 5 must be added to previous losses that accrue to each rule. The total expected loss for R_5 is greater than that for R_3 if and only if E(Cost) for R_5 is greater than that for R_3. The same holds true for R_6 compared with R_3. R_7 is preferable to R_4 if E(Cost) for R_7 is less than E(Cost) for R_4, and the same holds true for R_8 compared with R_4. The decision as to whether to consider sequential tests versus simultaneous tests is based purely on costs. Thus if $K_{12} = K_{21} = K_{(12)}$ or more particularly if $K_{ij} = K_{(12)} = K_1 + K_2$, then R_5 and R_6 will be preferable to R_3, and R_7 and R_8 will be preferable to R_4.

This whole discussion has assumed all of the parameters' values were known. Inevitably they are not and of course appropriate samples are required for their proper estimation whether frequentist or Bayesian. For some discussion of the latter issue, see Gastwirth et al. (1991) and Johnson and Gastwirth (1991). One also can foresee a situation of more than two binary

tests for a condition as is already true for AIDS. There is a more exacting test which is now an order of magnitude more expensive than the Western Blot, i.e. polymerase chain reaction test. Although in principle it can be handled in this manner, the number of potential rules increases exponentially so that methods for shortening the search for an optimal rule are also needed. In summary, there are many important and interesting problems here where biostatisticians can make enormous contributions to the optimal implementation of public policy in the health area.

Biostatisticians should not let other professionals - economists, risk analysts, etc. run away with this one because if they don't do it right it will be called a statistician's trick anyway.

3.2. The Issue of Interim Analysis in Clinical Trials

There are several issues that seem to have plagued frequentist Biostatisticians in the conduct and analysis of clinical trials. They are the sidedness of a test (one or two), multiple comparisons, and interim analyses. I will only address the latter and after a brief introduction confine it to a specific issue which interested me. For those interested in other aspects of interim analysis as well as this one, the recent review by Jennison and Turnbull (1990) should serve.

Briefly, given that a trial is planned to be analysed for a given number of subjects, how should making interim analyses during the course of the trial affect the final analysis. There are those who feel that inferences drawn should depend on the stopping rule, despite the Likelihood Principle. On the other hand they would also like to make interim analyses so that an unpromising trial could be abandoned early on, or at any rate, to be able to decide whether a trial is promising enough to continue. There are methods

which would permit control of type I and type II errors if interim analyses are made at preplanned particular sample sizes in a sequential trial, but they require that differences be appreciably larger to maintain the requisite Type I error compared to a trial where no interim analyses were to be made. Coupled with this is the fact that it is not always convenient to conduct interim analyses at the preordained sample sizes in a trial. Hence other methods have been attempted that permit unplanned interim analyses at arbitrary times but yield at best a very conservative test which makes the detection of significant differences even more difficult to assert.

Of course, the Bayesian statistician, who does not pay attention to a stopping rule that does not affect the likelihood, has no such problems. However, an important trial, especially one that is supposed to test the effectiveness of a new agent, should generally be carried out for at least some predetermined sample size if the scientific public or general public or a regulatory agency is to be convinced of the conclusions drawn. This assumes that all other things are equal. And even for a Bayesian this can lead to an interim analysis as to whether to continue a trial until its specified term. Also, it has been recognized even by those who might prefer a frequentist analysis that some Bayesian predictive input could be helpful in such a situation, Choi and Pepple (1989), Choi et al. (1985), Spiegelhalter et al. (1986).

We illustrate the idea that has been proposed by these statisticians in a very simple case. Let $X_1,...,X_{N+M}$ be i.i.d. $N(\theta,1)$ and require a test of the following hypothesis,

$$H_o: \theta \leq \theta_o \text{ vs. } H_1: \theta > \theta_o.$$

For test H_o vs. H_1 at level α, reject H_o if

$$\sqrt{N+M}\ (\bar{x}_{N+M}-\theta_o) > z_\alpha \tag{3.5}$$

where $\alpha = 1 - \Phi(z_\alpha)$, and $\Phi(\bullet)$ is the standard normal distribution function. Suppose we stopped after N observations and wanted to calculate the probability of achieving the above.

A syncretic approach has been suggested and developed in the previously mentioned papers which apply Bayesian predictive ideas towards the solution of this problem. It is assumed that the prior for θ is constant to conform as closely as possible to a frequentist analysis. After N observations are in hand, this results in a posterior distribution for θ as $N\left(\bar{x}_N, \frac{1}{N}\right)$.

Now we compute the probability of the aforementioned event

$$\Pr\left[(N+M)^{1/2}\left[\frac{N\bar{x}_N + M\bar{X}_M}{N+M} - \theta_o \right] > z_\alpha \right] = P_\alpha \tag{3.6}$$

where $\bar{x}_N$ is fixed and the future $\bar{X}_M$ is random so that the predictive distribution of $\bar{X}_M$ is easily obtained as $N\left(\bar{x}_N, \frac{1}{N} + \frac{1}{M}\right)$. Hence

$$P_\alpha = \Pr[N\bar{x}_N + M\bar{X}_M - \theta_o(N+M) \geq z_\alpha\sqrt{N+M}] \tag{3.7a}$$

or

$$P_\alpha = \Pr[M(\bar{X}_M - \bar{x}_N) + (M+N)(\bar{x}_N - \theta_o) \geq z_\alpha\sqrt{N+M}]. \tag{3.7b}$$

On regrouping terms in (3.7),

$$P_\alpha = \Pr\left[\frac{\bar{X}_M - \bar{x}_N}{\sqrt{\frac{1}{M}+\frac{1}{N}}} \geq \frac{z_\alpha\sqrt{N+M} - (M+N)(\bar{x}_N - \theta_o)}{M\sqrt{\frac{1}{M}+\frac{1}{N}}}\right] \tag{3.8a}$$

or

$$P_\alpha = \Pr\left[Z \geq \frac{z_\alpha\sqrt{N+M} - (M+N)(\bar{x}_N - \theta_o)}{M\sqrt{\frac{M+N}{MN}}}\right] \tag{3.8b}$$

where Z is N(0,1). Further

$$P_\alpha = \Pr\left[Z \geq \sqrt{\frac{N}{M}}\left[z_\alpha - (M+N)^{1/2}(x_N - \theta_o)\right]\right] \tag{3.9a}$$

$$P_\alpha = 1 - \Phi\left(\left(\frac{N}{M}\right)^{1/2}[z_\alpha - (M+N)^{1/2}(\bar{x}_N - \theta_o)]\right). \tag{3.9b}$$

This then is the probability that if the trial were continued for an additional M observations, H_o would be rejected at level α. Small values of P_α would discourage the continuation of the trial while large values would encourage it. But now consider the following easily established result,

$$\lim_{M\to\infty} P_\alpha = 1 - \Phi(-\sqrt{N}(\bar{x}_N - \theta_o))$$

$$= \Phi[\sqrt{N}(\bar{x}_N - \theta_o)] = 1 - P \tag{3.10a}$$

$$P = Pr[Z \geq \sqrt{N}(\bar{x}_N - \theta_0)], \tag{3.10b}$$

which is independent of α.

This implies that if one continued the trial indefinitely, the predictive probability of rejecting H_0 approaches 1-P irrespective of α. Note that this is a Bayesian interpretation of 1-P that naive students and some investigators often make with regard to significance tests; note also that teachers of frequentist statistics strive mightily to disabuse students of this flawed interpretation. Therefore I believe that upon reflection, neither a strict Bayesian nor Frequentist could accept such a result. On the one hand this result does not have an acceptable frequentist interpretation and on the other hand this is not the kind of test a Bayesian would apply. One perhaps needs to be careful of mixing metaphors.

A Bayesian approach in this situation would reject H_0, say, if the posterior probability

$$Pr[\theta > \theta_0 \mid x_1,\ldots,x_{N+M}] \geq p \tag{3.11}$$

assuming a prior $\pi(\theta)$ for θ. Hence, after N observations one would calculate the predictive probability of the above event assuming $x_1,\ldots,x_N$ have been observed and future observables $X_{N+1},\ldots,X_{N+M}$ are random. In this example if the previous prior for θ is used, then a posteriori

$$\theta \sim N\left(\bar{x}, \frac{1}{N+M}\right)$$

where

$$(N+M)\bar{x} = N\bar{x}_N + M\bar{x}_M$$

then H_o is rejected if

$$\Pr[\sqrt{N+M}\,(\theta-\bar{x}) > \sqrt{N+M}\,(\theta_o-\bar{x})] \geq p \tag{3.12a}$$

or

$$1 - \Phi(\sqrt{N+M}\,(\theta_o-\bar{x})) \geq p. \tag{3.12b}$$

Now stopping at N, we need to find the predictive probability of the above event i.e.

$$P_p = \Pr\left[1 - \Phi\left(\sqrt{N+M}\left(\theta_o - \frac{N\bar{x}_N + M\bar{X}_M}{N+M}\right) \geq p\right)\right]$$

$$= \Pr\left[(X_M - x_N) \geq \frac{(N+M)(\theta_o-\bar{x}_N)}{M} - \frac{\sqrt{N+M}\,\Phi^{-1}(1-p)}{M}\right]$$

$$= \Pr\left[\frac{X_M - x_N}{\sqrt{\frac{1}{M}+\frac{1}{N}}} \geq \frac{(N+M)(\theta_o-\bar{x}_N)}{M\sqrt{\frac{1}{M}+\frac{1}{N}}} - \frac{\sqrt{N+M}\,\Phi^{-1}(1-p)}{M\sqrt{\frac{1}{M}+\frac{1}{N}}}\right],$$

and finally we obtain

$$P_p = 1 - \Phi\left(\left(\frac{N}{M}\right)^{1/2}[(\theta_o-\bar{x}_N)(N+M)^{1/2} - \Phi^{-1}(1-p)]\right). \tag{3.13}$$

Now if the trial were contemplated to be continued indefinitely,

$$\lim_{M\to\infty} P_p = 1 - \Phi\left((\theta_o - \bar{x}_N)\sqrt{N}\right) = Pr\left[\theta > \theta_o \mid x_1,\ldots,x_N\right] \qquad (3.14)$$

which does not depend on p and is obviously the posterior probability given N observations. This is perfectly sensible as the best prediction of what would occur if one were to continue sampling indefinitely. It appears that if frequentists start down the slippery slope of Bayesianism they might as well slide all the way.

More useful Bayesian applications to interim analysis are given in Geisser (1991).

3.3. Statistical Regulation of Chronic Diseases

Although the control or regulation of "abnormal" physiological or behavioral variables or symptoms associated with a condition or chronic ailment does not have quite the same appeal that therapeutic cures or surgical intervention possess, it is probably the most pervasive of medical practice and deserves greater attention than it has previously been accorded by Biostatisticians.

An important goal in a variety of chronic conditions such as diabetes, arthritis, high blood pressure and high cholesterol levels is to regulate or control a physiological variable by an appropriate administration of a therapeutic agent. For example, severe diabetics need to control their glucose level by carefully regulated injections of insulin one or more times a day. The amount infused which is controllable will depend on the current glucose

level and other both controlled or uncontrolled covariates. What is required is a modeled functional relationship of the form

$$E(Y_{t+1}) = f(y_t, v_t, x_t, \theta), \tag{3.15}$$

Y_{t+1} = response to be controlled at t+1

y_t = measured response at time t

v_t = amount of drug to be administered at time t

x_t = value of a set of covariates at time t

θ = set of unknown parameters in the functional relationship.

For the most recent frequentist way of handling autoregressive models using squared error for prediction see Lai and Zhu (1991).

Models of this sort have been used for engineering problems and they also appear in the economics literature e.g. Zellner (1971), but infrequently in the biostatistical literature, especially in regard to control of chronic conditions. And until "cures" for these are found, it is important that the course of these chronic conditions be properly regulated.

Assuming for a diabetic that at time t we can specify the amount of insulin infused then we would want to do so to ensure that the response Y_{t+1} will be as close as possible to some given value say y_o or with maximal probability to be in an interval (y_o+a, y_o+b) for $a < b$.

We could perhaps consider the overly simplified linear model for the amount of glucose in a diabetic

$$Y_{t+1} = \alpha y_t + \beta x_t + \gamma v_t + e_{t+1}$$

where e_{t+1} is $N(0,\sigma^2)$ and independent for each t, y_t represents the amount of glucose, x_t the caloric intake and v_t the amount of insulin to be infused to regulate the value Y_{t+1} in some interval. I suspect that even this very simple model, with possibly some elaboration, could be adequate for regulation purposes in many situations. In general this type of problem is best served by obtaining a probability distribution of values for Y_{t+1}, using a Bayesian approach based on having observed previous values (y_j,v_j,x_j) $j=1,...,t$ and some prior distribution for the parameters (α,β,γ). Hence for a value as close as possible to y_o we could use the predictive squared error

$$\hat{v}_t = \min_{v_t} E(Y_{t+1}-y_o)^2 = \min_{v_t} [V(Y_{t+1}) + (E(Y_{t+1}) - y_o)^2]$$

or that value of v_t that maximizes the predictive probability function of Y_{t+1} evaluated at $Y_{t+1} = y_o$ using notation $a^{(k)} = (a_1,...,a_k)$,

$$\max_{v_t} f_{Y_{t+1}}(y_o \mid y^{(t)},x^{(t)},v^{(t-1)},v_t).$$

Sometimes other loss functions that are asymmetric are appropriate e.g. if a diabetic may need to be more wary of going into insulin shock (i.e. too little glucose (hypoglycemia)) as opposed to acidosis (i.e. too much glucose (hyperglycemia)), or vice versa an asymmetric loss function may serve. If an interval is required we then search for that v_t which maximizes

$$Pr[y_o+a \le Y_{t+1} \le y_o+b].$$

The values a and b would yield a symmetric interval about y_0 if a = -b. One could also set a and b to values that to some degree stress the importance of being above or below y_0 or use a linex loss function Zellner (1986).

The most difficult issue, of course, is the formulation of an appropriate model that adequately reflects the process. Once this is done and sufficient data has been obtained, the problem of modeling the prior distribution will be relatively unimportant and one can develop numerical methods to obtain the values for regulating the response. In fact, for any chronic condition, one looks forward to the time when a hand-held programmable calculator can be used to compute the approximate therapeutic dose given the current and past data, or even more elaborately, a computer-controlled infusion device.

In conclusion, I hope I have pointed out a few of the interesting statistical issues and opportunities involved in forensics, medicine and public affairs, not to mention a few idiosyncratic peeves.

References

Aitken, M. (1991). Evidence and the posterior Bayes factor. Unpublished manuscript of a talk delivered at a plenary session of the Israeli Statistical Association, May 16, 1991.

Ambrose, S.E. (1991). Ike and the disappearing atrocities. New York Times Book Review Section, February 24, 1.

Bacque, J. (1989). Other Losses: An Investigation into the Mass Deaths of German Prisoners at the Hands of the French and Americans After World War II, Stoddart, Toronto.

Berry, D.A. (1991). Inferences using DNA profiling in forensic identification and paternity cases. Statistical Science, 6, 2, 175-180.

Budowle, B. et al. (1991). Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. American Journal of Human Genetics 48, 841-855.

Choi, S.C. and Pepple, P.A. (1989). Monitoring clinical trials based on predictive probability of significance. Biometrics, 45, 317-323.

Choi, S.C., Smith, P.J. and Becker, D.P. (1985). Early decision in clinical trials when treatment differences are small. Controlled Clinical Trials, 6, 280-288.

Cowdry, A.E. (1990). Review of "Other Losses". Canadian Bulletin of Medical History, 7, 187-91.

Gastwirth, J.L., Johnson, W. and Reneau, D.M. (1991). Bayesian analysis of screening data: Application to AIDS in blood donors. Canadian Journal of Statistics, 19, 2, 135-150.

Geisser, S. (1991). On the curtailment of sampling. Canadian Journal of Statistics (in press).

Geisser, S. and Johnson, W.O. (1991). Optimal administration of dual screening tests for detecting a characteristic with special reference to low prevalence diseases. Biometrics (in press).

Jennison, C. and Turnbull, B.W. (1990). Interim monitoring of medical trials. Statistical Science 5, 3, 299-317.

Johnson, W.O. and Gastwirth, J.L. (1991). Bayesian inference for medical screening tests: Approximations useful for the analysis of AIDS data. Journal of the Royal Statistical Society B, 53, 2, 427-440.

Lai, T.L. and Zhu, G. (1991). Adaptive prediction in non-linear autoregressive models and control systems. Statistica Sinica, 1, 2, 309-334.

Lane, D.A. (1989). Subjective probability and causal assessment. Applied Stochastic Models and Data Analysis, 5, 53-76.

Lane, D.A. (1991). Causal inference from case reports. Probability in Biology and Medicine, eds. M. di Bacco and G. Coletti, Springer (in press).

Speigelhalter, D.J., Freedman, D.S. , and Blackburn, P.R. (1986). Monitoring clinical trials: Conditional or predictive power? Controlled Clinical Trials 7, 8-17.

Zellner, A. (1971). An Introduction to Bayesian Inference Econometrics, John Wiley: New York.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. Journal of the American Statistical Association, 81, 446-451.